ISSN No. 0976-5697

Volume 11, Special Issue I, May 2020

International Journal of Advanced Research in Computer Science

CONFERENCE PAPER

Available Online at www.ijarcs.info

Kannada Kali

Lalitha L A
Dept of C&IT,REVA University
Bengaluru, India
gouthamjason@gmail.com

Goutham Jason,
Dept of C&IT, REVA University
Bengaluru, India
gouthamjason@gmail.com

Harshith S
Dept of C&IT, REVA University
Bengaluru, India
harshith1.sheshan@gmail.com

Kenneth Jones
Dept of C&IT, REVA University
Bengaluru, India
kennethjones1737@gmail.com

Kevin George Thomas
Dept of C&IT,REVA University
Bengaluru, India
kevingt2007@gmail.com

Abstract:This project is an Application to make a user understand and learn the basics of a language. Kannada Kali focuses on the 4 main aspects of language learning: Listening, Reading, Speaking and Writing. While a child learns, detecting real word error is a really difficult task and requires advanced statistical processing, Data Mining and Natural Language Processing (NLP) techniques which we have implemented in this project.

Keywords: Natural Language Processing, Convolutional Neural Networks, Sequence Matcher, Spaced Repetition, Django

I. INTRODUCTION

Natural language processing (NLP) is a domain of artificial intelligence where a model/computer or any processing data unit can comprehend, understand, analyze and derive a certain meaning/outcome of the data which is used in a smart and useful way. The application of NLP allows developers to organise and create a structure of that knowledge which can be utilized into performing a goal-oriented task such as Segmentation, Sentiment Analysis, Named Entity Recognition etc.

When compared to other species, natural language is one of the primary unique selling propositions of the human mind. NLP, a major buzzword in today's tech discussion, deals with how computers can understand and generate language. The rise of NLP in the past decades is backed by a couple of global developments: the universal hype around AI, exponential advances in the field of Deep Learning and an ever-increasing quantity of available text data. But what is the substance behind the buzz? In fact, NLP is a highly complex, interdisciplinary field that is constantly supplied by high-quality fundamental research in linguistics, math and computer science.

Our project here focuses on a certain aspect, aimed at bridging the learning gap in education. According to the Annual Status of Education Report (ASER) 2018, of the majority of students enrolled in a lower grade class such as 5, approximately only 50% could read a grade 2 textbook confidently. This lack of reading ability undoubtedly played a large role in further education and finally to harness the child's true potential. Although there can be various other parameters simple and complex, the lag in learning revolves around poor infrastructure, mediocre quality of learning material etc.

What differs us from other products in the market would be the addition of Spaced Repetition. Spaced repetition is a proven theory-practical approach and a learning technique that uses flashcards. The introduction of newer and more difficult data is presented while previous and older are much lesser. This is done to exploit a feature known as the 'Psychological spacing effect'. It is proven that this technique increases the learning rate exponentially and used in a certain scenario where a learner is necessary to acquire large data over a shorter period of time and indefinitely.

It is, therefore, well suited for the problem of vocabulary acquisition in the course of second-language learning. This completely ignores human intervention and focuses neat and simple methodological teaching, systematically.

We are in the age of computers in which the local languages have necessarily to be used to keep them from the onslaught of English and, in consequence, from their extinction. Owing to the present education system and the indifferent attitude of the users towards their own language, and lack of interest in them in learning the local language and the resources, we come across careless writings in all fields of communication. To avoid this and to help using the local languages on computers, it is absolutely necessary that we should have a 'Learning Medium' which is required to identify and correct the mistakes committed by users. But, so far, users of Kannada Language have not got a fairly accurate medium. This approach also enables us to design and implement techniques in the domain of Artificial Intelligence using Natural Language Processing which is a learning opportunity in itself.

II. LITERATURE SURVEY

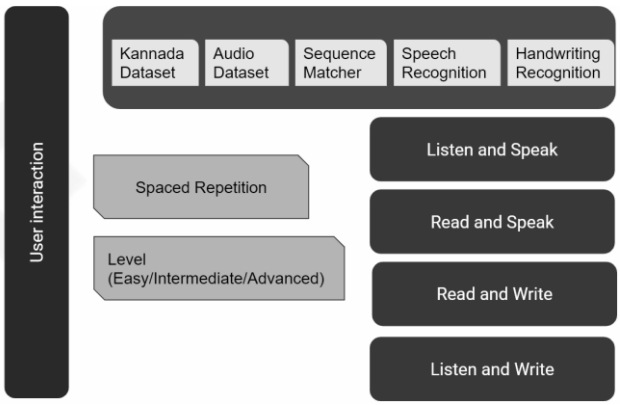
Mobile applications being used to learn a foreign language or to learn a second language has attracted many users. Revenue generated by English language learning products crossed $1.8 billion in 2013 and reached $3.1 billion around 2018, with a compounded growth rate of 11.1%. Language learning apps like DuoLingo have expanded with over 90 million downloads. Demand for language learning apps will further increase exponentially as the accessibility of mobile computing devices increases.[1]

The scripts in Indian languages have their essence drawn from the ancient Brahmi Script. The basic units are referred to by Akshara's. Depending on the context, an Akshara can have a complex shape with the position of consonant and vowel being placed on top, below, before, after or sometimes surrounding the pivotal symbol. Thus, to complete an Akshara, a set of semi-full or half-forms have to be rendered, which are represented using shapes referred to as glyphs.[2]

Speech Recognition is the process of converting a speech signal to a sequence of words by means of an Algorithm implemented as a computer program. Speech to text systems can be classified into two types, such as speaker-dependent and speaker-independent systems. There are various methods of feature extractions. In recent researches, many feature extraction techniques are commonly used such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Cepstral Analysis and Mel-frequency cepstral (MFCCs), Kernel-based feature extraction based approach, Wavelet Transform and spectral subtraction [3]

A text to speech module is destined to convert the given text to corresponding waveform in which it is spoken. The main components of a text to speech system are text processing and generation. Text processing handles the task of processing the input text to produce phonemic units. The speech generation handles the realization of these units. For the process to be natural it is important that the text be processed to appropriate phonemic units which correspond to the input text.[4]

According to Peter Norvig's approach, we generate all the possibilities of misspelled words using three main kinds of modification which are add, remove and replace. When considering misspelled words there exist a lot of possible combinations. We are left with a task to choose from these huge number of misspelled forms of which we consider only the proper meaningful words as per the dictionary of the language we are considering which narrows down our possibilities. We then rank the possible words according to the frequency of occurrence in our context to consider relevancy as a factor. Then we choose the best possible match with the highest possibility and relevance.[5]



Google Bolo Application which is still in beta version has been a great effort from google to teach kids to read local languages. The application has no support for Kannada yet and has the ability to only teach the user to read. We have taken inspiration from the application and also figured where the application lacks and will try to improve upon it in our application.[6]

III. PROPOSED SYSTEM

A. System Architecture

Figure 1. Architecture of the system.

B. Methodology

The initial requirement of this project is a massive vocabulary, which we have created using Kannada textbooks and ordered the words based on important criteria such as complexity, frequency, etc. Each word is exposed to the child to read, listen, write and pronounce.

We have also created an audio file of every word for the listen to related modules. We utilize the python speech recognition package to recognize the word and then use a sequence matcher to find the similarity quotient

and check if it satisfies a high threshold of 0.9 to accept the response. For the purpose of teaching, we have opted to choose two methodologies/techniques which are spaced repetition and using Levels of words to determine the sequence of progress.

For the write related modules, we trained our model on the hand-drawn Kannada characters in the Chars74K dataset using Convolutional Neural Networks. We perform character segmentation before the handwriting recognition to efficiently detect the ottaksharas present in the word.

IV. ALGORITHMS

Accurately classifying the correct answer is an important part of maintaining user interaction. Many algorithms are utilized for comparison and understanding of user input.

A. Sequence Matcher

Sequence Matcher is a flexible class for comparing pairs of sequences of any type, so long as the sequence elements are hashable. The idea is to find the longest contiguous matching subsequence that contains no "junk" elements. The same idea is then applied recursively to the pieces of the sequences to the left and to the right of the matching subsequence. This does not yield minimal edit sequences but does tend to yield matches that "look right".

B. Speech Recognition Library

The Speech Recognition library utilizes algorithms like Perceptual Linear Prediction (PLP) features, Viterbi search, deep neural networks, discrimination training, and Weighted Finite-State Transducers (WFST) framework to accurately recognize speech.

The SpeechRecognition library acts as a wrapper for several popular speech APIs and is thus extremely flexible. One of these—the Google Web Speech API—supports a default API key that is hard-coded into the SpeechRecognition library. The library also has support for the Kannada language and is easily able to recognize audio from the user.

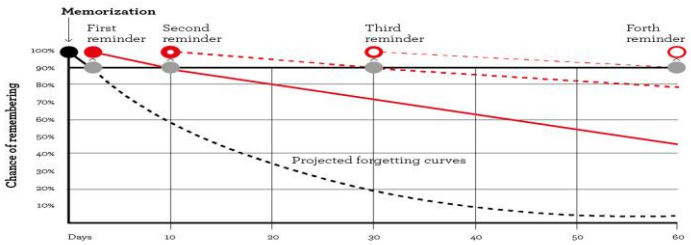
C. Convolutional Neural Networks

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNetshave the ability to learn these filters/characteristics. Using this algorithm, we are able to understand the user's handwriting.

```
scoreuser():
    if( user passes test case):
        score++
    else:
    score--
Frequency[] #we use frequency as a counter to find how often a word repeats
Pass[] #pass is another counter to stop the loop if less than 3 words are left
Goal score: 3 points
while(till all words reach goal score):
    if(score[i]<Goal score && frequency[i]>3):
        repeat word
        scoreuser()
    else if(score<3 && frequency[i]<3 && pass[i]<3):
        pass[i]++
        continue() #skip the word
    else:
        repeat word
      scoreuser()}
```

D. Spaced Repetiton

We designed a spaced repetition model based on the method of reinforcement learning and considering scoring words based on the performance of the user. The scheme for scoring that we used is as follows:



V. IMPLEMENTATION AND RESULTS

A. Spaced Repetiton

We designed a sequence matching technique base on Peter Novig's idea for English and made the same algorithm language free and got a performace metric of 86.8 % accuracy.

```
[22] for i in range(1000):
         if((df.loc[i,'Kannada'])==(df.loc[lens[i],'Kannada'])):
             count=count+1
```

```
print("Accuracy =",count*100/1000,"%")
```

```
Accuracy = 86.8 %
```

Figure 2. Performace of Sequence Matcher

B. User Interface

Our algorithms and many techniques are presented to the user through a web application. We utilize Django as our web framework to be able to run python scripts for the web application. Django is also connected to our PostgreSQL database containing the sequence of words and also the user's completion rate.

C. Recognizing User Audio

User audio is recognized using the speech recognition library available in python. It is one of the most efficient libraries that exist for audio recognition and has strong support for the Kannada language. We wanted the module to produce a positive result matching the context in spite of the misinterpretation of a few syllables. Hence we use a sequence matcher to compare the output received with the required output. If the two words have a

match of 0.9 or higher, we consider the output to be successful, else we show the mistake to the user and ask him to try again.

D. Recognizing User Handwriting

The most difficult part of the application is to accurately identify the user's handwriting. Many handwriting models exist for popular languages like English but there is very little support for the Kannada language. Thanks to the Chars74K image dataset [8], we are able to create a Convolutional Neural Network (CNN) model to be able to accurately recognize handwritten characters. Since none of the characters in Kannada are connected, we can perform character segmentation to accurately divide the word into characters along with it'sottaksharas.

E. Utilizing Spaced Repetition

Some words are more difficult than others so we store the number of attempts taken by a user to achieve the required output. Using this data, we are able to assign a strength of the word to the user so that this word may be repeated again to help the user perfect the word thereby increasing its strength.

Figure 3. Effect of Spaced Repetition

VI. CONCLUSION AND FUTURE ENHANCEMENTS

In this work, we've explained the basic architecture and the modules used in our project. Our application helps a student of any age learn the four basic aspects of language learning which are listening, speaking, reading and writing for Kannada. The main intent for this project was to take it up to a social cause and do something for the less fortunate people in the society who cannot afford the form of education most of us enjoy as a privilege. Currently, the project is capable of being extended to other regional languages if the dataset is interchanged with the user's choice of language.

If our project meets our expected target and successfully helps a student learn Kannada in the span of 4-6 weeks. We intend on involving the local govt institutions to test our prototype. The project will be kept as an open source in the PyPI library and we ask others to use our work for reference. The initial results of our model can be used to further improve the performance of our model. We also plan on using audio filters such as de-essing, noise-reduction, compression, and a low-pass filter in order to enhance the attributes in our audio files such as pronunciations, background noise elimination and so on.

VII. REFERENCES

[1] Catherine Regina Heil, Jason S. Wu, Joey J. Lee, Torben Schmidt (2016). "A review of mobile language learning applications: trends, challenges and opportunities", The EUROCALL Review, Volume 24, No. 2, September 2016

[2] Prahallad L., Prahallad K., and Ganapathiraju M., "A simple approach for building transliteration editors for Indian languages," Journal of Zhejiang University Science, vol. 6A, no. 11, pp. 1354–1361,2005

[3] Anand Arokia Raj, Tanuja Sarkar, Satish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad1, Alan W Black, "Text Processing for Text-to-Speech System in Indian Languages." K. Elissa, "Title of paper if known," unpublished.

[4] Su Myat Mon, Hla Myo Tun, "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM) "

[5] Peter Norvig, "Artificial Intelligence A Modern Approach-Prentice Hall (2010)"

[6] Google Bolo application, Google LLC press release March 2019

[7] The Chars74k image dataset - http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/